UNIVERSIDAD CARLOS III DE MADRID working papers

Working Paper 01-64
Business Economics Series 13
November 2001

Departamento de Economía de la Empresa
Universidad Carlos III de Madrid
Calle Madrid, 126
28903 Getafe (Spain)
Fax (34-91) 6249608

Robust Logistic Regression for Insurance Risk Classification

Esteban Flores * and José Garrido **

Abstract

Risk classification is an important part of the actuarial process in Insurance companies. It allows for the underwriting of the best risks, through an appropriate choice of classification variables, and helps set fair premiums in rate-making.
Logistic regression is one of the sophisticated statistical methods used by the banking industry to select credit rating variables. Extending the method to insurance risk classification seems natural. But Insurance risks are usually classified in a larger number of classes than good and bad, as is usually the case in credit rating.
Here we consider a model generalization to extend the use of logistic regression to insurance risk classification. Since insurance data presents catastrophic losses and heavy tail claim distributions, robust estimation will be important. A new robust regression estimator for the logistic model, both in the binary and multinomial response cases, is proposed. Its asymptotic properties are also studied.

Key words: Minimum distance estimation; quadratic distance; logistic regression; robustness; asymptotic normality; multinomial logistic regression; risk classification.

*Esteban Flores, Department of Mathematics and Statistics, Concordia University, Montreal, Qc H4B 1R6, Canada, e-mail: ores@alcor.concordia.ca and University of Talca, Talca, Chile.

**José Garrido, Department of Mathematics and Statistics, Concordia University, Montreal, Qc H4B 1R6, Canada, e-mail: garrido@vax2.concordia.ca and Department of Business Administration, University Carlos III of Madrid, Colmenarejo, Madrid, 28270 Spain.

1 Introduction

A minimum distance method based on a quadratic distance was introduced by Luong and Thompson (1987). Following the same idea a minimum quadratic distance estimator (QDE) was defined by Luong (1991) for the simple linear regression model. An extension to multiple linear regression was studied by Luong and Garrido (1992), where the asymptotic properties of this QDE were derived. They show that the QDE is fully efficient, for special choices of odd functions h_i in the distance definition, and robust for other appropriate choices of h_i.

In Section 2, a QDE is defined for the logistic regression model. The asymptotic properties of this QDE are derived, where consistency, asymptotic normality and robustness properties are established.

In Section 3, a new robust QDE for the multinomial logistic regression model (QDM) is proposed. The asymptotic normality property is established using the approach developed in previous sections.

2 Robust Quadratic Distance Estimators for Logistic Regression

Let $\mathbf{x}_i^T = (x_{i1}, \ldots, x_{ip}) \in \mathbb{R}^p$ be a vector of p (discrete or continuous) explanatory variables and $\mathbf{X} = (\mathbf{X}_1^T, \ldots, \mathbf{X}_N^T)^T$ be a $N \times p$ design matrix of rank $p \leq N$, with $\mathbf{x}_i \neq 0$, for $i = 1, \ldots, N$. Denote by $n^* = \lfloor \frac{N}{2} \rfloor + \lfloor \frac{p}{2} \rfloor$, where $\lfloor z \rfloor$ stands for the largest integer less than or equal to $z \in \mathbb{R}$.

Consider a logistic regression model for binary responses and N (not necessarily independent) random variables Y_i, which have a binomial distribution with index n_i and probability $\pi(\mathbf{x}_i)$. These are denoted by $Y_i \sim$ Binomial$(n_i, \pi(\mathbf{x}_i))$, where n_i is a known positive integer, $\pi(\mathbf{x}_i) = \frac{\exp(\mathbf{x}_i^T \boldsymbol{\beta})}{1+\exp(\mathbf{x}_i^T \boldsymbol{\beta})}$ and $\boldsymbol{\beta}$ is a vector of p unknown parameters.

As in Christmann (1994) the relative frequencies P_i can defined as follows.

Definition 2.1. Let y_i be observations from $Y_i \sim \text{Bi}(n_i, \pi(\mathbf{x}_i))$, then the relative frequencies P_i are defined, for $i = 1, \ldots, N$, as

$$P_i = \begin{cases} \frac{1}{2n_i} & \text{if } Y_i = 0 \quad , \\ \frac{Y_i}{n_i} & \text{if } 1 \leq Y_i \leq n_i - 1 \quad , \\ 1 - \frac{1}{2n_i} & \text{if } Y_i = n_i \quad . \end{cases} \tag{1}$$

Assumptions: Under the above definitions it is assumed that

(a) there exist $\pi(\mathbf{x}_i) \in (0,1)$, for $1 \le i \le N$, such that if $\min_{1\le i\le N} n_i \longrightarrow \infty$

$$(P_1, \ldots, P_N) \longrightarrow (\pi(\mathbf{x}_1), \ldots, \pi(\mathbf{x}_N)) \ , \quad \text{almost surely} \ , \tag{2}$$

(b) there exists exactly one vector $\boldsymbol{\beta}^* \in \mathbb{R}^p$ such that, for all $\boldsymbol{\beta} \neq \boldsymbol{\beta}^*$,

$$\left| \left\{ i; \pi(\mathbf{x}_i) = \frac{e^{\mathbf{x}_i^T \boldsymbol{\beta}^*}}{1 + e^{\mathbf{x}_i^T \boldsymbol{\beta}^*}} \right\} \right| \ge n^* > \left| \left\{ i; \pi(\mathbf{x}_i) = \frac{e^{\mathbf{x}_i^T \boldsymbol{\beta}}}{1 + e^{\mathbf{x}_i^T \boldsymbol{\beta}}} \right\} \right| \quad . \tag{3}$$

The strong law of large numbers guarantees the validity of (2) for the logistic regression model. Then (3) holds by definition of $\pi(\mathbf{x}_i) = \frac{\exp(\mathbf{x}_i^T \boldsymbol{\beta}^*)}{1+\exp(\mathbf{x}_i^T \boldsymbol{\beta}^*)}$ and $\text{rank}\,(\mathbf{X}) = p$.

In what follows it is assumed that all values of n_i are reasonably large, in the sense that the results are asymptotic for $n_\cdot = \sum_{i=1}^N n_i \to \infty$ such that $\frac{n_i}{n_\cdot} \to c_i \in (0,1)$, but N and p remain fixed.

Theorem 2.1. If the logistic model holds true and n_i is large, then the empirical logit transform $\ln\left(\frac{P_i}{1-P_i}\right)$ is approximately normally distributed with mean $\mathbf{x}_i^T \boldsymbol{\beta}$ and variance $\{n_i \pi(\mathbf{x}_i)(1-\pi(\mathbf{x}_i))\}^{-1}$, that is

$$\ln\left(\frac{P_i}{1-P_i}\right) \ \approx \ N(\mathbf{x}_i^T \boldsymbol{\beta}, \{n_i \pi(\mathbf{x}_i)(1-\pi(\mathbf{x}_i))\}^{-1}) \ , \quad \text{for } i = 1, \ldots, N \ .$$

Proof. See Appendix A.

Definition 2.2. (a) Let $\tilde{\mathbf{X}}^T = (\tilde{\mathbf{X}}_1^T, \ldots, \tilde{\mathbf{X}}_N^T)$ be the $N \times p$ matrix of transformed (discrete or continuous) explanatory variables $X_1, \ldots, X_p$, with $\tilde{\mathbf{X}}_i^T = v_i \mathbf{x}_i^T$, where $v_i = \{n_i P_i (1-P_i)\}^{\frac{1}{2}}$, for $i = 1, \ldots, N$.

(b) Let $\tilde{\mathbf{Y}}^T = (\tilde{Y}_1, \ldots, \tilde{Y}_N)$ be the $N \times 1$ vector of the empirical logit transform, where $\tilde{Y}_i = v_i \ln\left(\frac{P_i}{1-P_i}\right)$, for $i = 1, \ldots, N$.

(c) By means of (a) and (b) for any $\boldsymbol{\beta} \in \mathbb{R}^P$, define the 'residual' as

$$\tilde{r}_i = \tilde{y}_i - \tilde{\mathbf{x}}_i^T \boldsymbol{\beta}, \quad \text{for } i = 1, \ldots, N \ . \tag{4}$$

Under the definition of residuals in (4), the logistic regression model can be considered as a particular case of the multiple linear model studied by Luong and Garrido (1992) in the context of quadratic distance estimation.

In what follows assume that the random errors

$$\tilde{r}_i = \tilde{y}_i - \tilde{\mathbf{x}}_i^T \boldsymbol{\beta}_0 \ , \quad \text{for } i = 1, \ldots , N \ ,$$

where $\boldsymbol{\beta}_0^T = (\beta_{01}, \ldots , \beta_{0p})$ is the vector of unknown parameters, are independent and identically distributed. Their common distribution function, F_0, is unknown (non-parametric model) but assumed to be absolutely continuous with a density function f_0, symmetric around zero. In fact, using Theorem 2.1 it is simple to check that the expected value and the index of skewness of the random errors are both equal to zero.

Define, for any $\boldsymbol{\beta} \in \mathbb{R}^p$

$$\hat{F}_j^{\boldsymbol{\beta}}(y) \ = \ \sum_{i=1}^{N} w_{ij} I(\tilde{y}_i - \tilde{\mathbf{x}}_i^T \boldsymbol{\beta} \leq y) \ , \quad \text{for } j = 1, \ldots , p \ , \tag{5}$$

where I denotes the indicator function and w_{ij} are known weights. Similarly, define

$$F_j^0(y) \ = \ \sum_{i=1}^{N} w_{ij} F_0(y) \ , \quad \text{for } j = 1, \ldots , p \ . \tag{6}$$

Note that $\hat{F}_j^{\boldsymbol{\beta}}$ are empirical processes based on the residuals in (4) and the known weights $w_{1j}, \ldots , w_{Nj}$, while F_j^0 are the corresponding theoretical distributions.

Also define for $j = 1, \ldots , p$

$$\begin{aligned} \mathbf{Z}_j^{\boldsymbol{\beta}} \ &= \ \left[\int_{-\infty}^{\infty} h_1(x) d\hat{F}_j^{\boldsymbol{\beta}}(x), \ldots , \int_{-\infty}^{\infty} h_k(x) d\hat{F}_j^{\boldsymbol{\beta}}(x) \right]^T \\ &= \ \left[\sum_{i=1}^{N} w_{ij} h_1(\tilde{y}_i - \tilde{\mathbf{x}}_i^T \boldsymbol{\beta}), \ldots , \sum_{i=1}^{N} w_{ij} h_k(\tilde{y}_i - \tilde{\mathbf{x}}_i^T \boldsymbol{\beta}) \right]^T \ , \\ \text{and} \quad \mathbf{Z}_j^0 \ &= \ \left[\int_{-\infty}^{\infty} h_1(x) dF_j^0(x), \ldots , \int_{-\infty}^{\infty} h_k(x) dF_j^0(x) \right]^T \ , \end{aligned}$$

where $h_1, \ldots , h_k$ is a fixed choice of odd functions, i.e. $h_i(x) = -h_i(-x)$, for $x \neq 0$ and $h_i(0) = 0$.

The minimum quadratic distance estimator (QDE) is the vector $\hat{\boldsymbol{\beta}}$ which minimizes the following sum of quadratic forms

$$d(\boldsymbol{\beta}) \;=\; (\mathbf{Z}_1^{\boldsymbol{\beta}} - \mathbf{Z}_1^0)^T \mathbf{Q} (\mathbf{Z}_1^{\boldsymbol{\beta}} - \mathbf{Z}_1^0) + \cdots + (\mathbf{Z}_p^{\boldsymbol{\beta}} - \mathbf{Z}_p^0)^T \mathbf{Q} (\mathbf{Z}_p^{\boldsymbol{\beta}} - \mathbf{Z}_p^0) \;, \qquad (7)$$

where $\mathbf{Q}$ denotes a $k \times k$ constant, symmetric, positive-definite matrix.

Furthermore, since $\mathbf{Z}_j^0 = 0$ for $j = 1, \ldots, p$, when h is odd, minimizing (7) with respect to $\boldsymbol{\beta}$, is reduced to minimizing

$$d(\boldsymbol{\beta}) \;=\; [\mathbf{Z}_1^{\boldsymbol{\beta}}]^T \mathbf{Q} \mathbf{Z}_1^{\boldsymbol{\beta}} + \cdots + [\mathbf{Z}_p^{\boldsymbol{\beta}}]^T \mathbf{Q} \mathbf{Z}_p^{\boldsymbol{\beta}} \;. \qquad (8)$$

Using Kronecker's product notation [see, Graham (1981)] and calling $\mathbf{Z}^{\boldsymbol{\beta}} = ([\mathbf{Z}_1^{\boldsymbol{\beta}}]^T, \ldots, [\mathbf{Z}_p^{\boldsymbol{\beta}}]^T)^T$, then (8) can be expressed more concisely as

$$d(\boldsymbol{\beta}) \;=\; [\mathbf{Z}^{\boldsymbol{\beta}}]^T (\mathbf{I}_p \otimes \mathbf{Q}) \mathbf{Z}^{\boldsymbol{\beta}} \;, \qquad (9)$$

where $\mathbf{I}_p$ denotes the identity matrix of order p. The QDE $\hat{\boldsymbol{\beta}}$ is the vector which minimizes (9) with respect to $\boldsymbol{\beta}$.

Example 1.

i	x_i^T		y_i	n_i
1	1	281.5	56	68
2	1	750.0	64	75
3	1	1375.0	54	67
4	1	2375.0	68	79
5	1	4000.0	46	56
6	1	6250.0	41	46
7	1	8750.0	33	42
8	1	12500.0	37	45
9	1	20000.0	46	53
10	1	37500.0	53	55
11	1	67500.0	66	70
12	1	90000.0	46	50
13	1	97500.0	83	93

Table 1: Fire data set

We illustrate the implementation of the QDE with a real data set given in Table 1. These fire claims are from a Danish portfolio of dwellings. For

each loss amount, the total floor space x_i (in square meters) of the insured dwelling is given; it is the only classification variable available here.

Different x_i values define classes $i = 1, \ldots, 13$. The corresponding number of claims in class i is denoted n_i. Then the portfolio is divided in two groups; losses less than 22,065 Danish Kroner (9 Dkr.$\cong$ 1 US$) belong to the first group, the total number of which is denoted y_i.

Using a simple logistic regression model we estimate the parameters by two methods: maximum likelihood (MLE) and our QDE (see Table 2). No outliers were detected using the pre-programmed function *lmsreg* (least median of squares robust regression) in S-Plus.

	MLE	QDE
$\hat{\beta}_0$	1.650744	1.650744
$\hat{\beta}_1$	9.106339E-6	8.795589E-6

Table 2: Estimation of parameters

The QDE was then obtained using the optimal weight matrix $\mathbf{W} = \tilde{\mathbf{X}}(\tilde{\mathbf{X}}^T\tilde{\mathbf{X}})^{-1}$. Since $\mathbf{Q} = \mathbf{\Sigma}^{-1}$ depends on the beta parameter values, we use the MLE as initial value to obtain $\mathbf{Q}$ iteratively. Finally, for the minimization of $d(\boldsymbol{\beta})$ in (9), we chose the following functions

$$h_1(x) = \begin{cases} 1 & \text{if} \quad x > 0 \\ 0 & \text{if} \quad x = 0 \\ -1 & \text{if} \quad x < 0 \end{cases} \qquad h_2(x) = \begin{cases} x & \text{if} \quad |x| \leq M \\ \text{sign}(x)M & \text{if} \quad |x| > M \end{cases} .$$

As expected, the estimators in Table 2 are essentially the same under the two methods when outliers are not present. Section 2.2 illustrates the effect that outliers have on both estimators.

2.1 Asymptotic Properties of the QD Estimator

This section establishes the consistency and asymptotic normality of the QDE. The derivation is based on the results of Luong and Garrido (1992) for multiple linear regression, adapted here to logistic regression.

Definition 2.3. Let $\mathbf{W}^T = (\mathbf{w}_1^T, \ldots, \mathbf{w}_N^T)$ be the $N \times p$ matrix of weights used in (5), where $\mathbf{w}_i^T = (w_{i1}, \ldots, w_{ip})$, for $i = 1, \ldots, N$.

Theorem 2.2. [**Consistency**] Consider the $N \times p$ matrix of weights $\mathbf{W}$ defined above. Let $\tilde{\mathbf{X}}$ be the $N \times p$ matrix given in Definition 2.2 (a), where $\mathbf{W}$ and $\tilde{\mathbf{X}}$ are assumed to have rank p. If the weights matrix $\mathbf{W}$ satisfies assumption $\boldsymbol{(a1)}$ in Appendix A, then the QDE $\hat{\boldsymbol{\beta}}$, obtained minimizing the function $d(\boldsymbol{\beta})$, is consistent.

Proof. Chebyshev's inequality and assumption $\boldsymbol{(a1)}$ give that $\mathbf{Z}^{\boldsymbol{\beta}_0} \xrightarrow{P} 0$, provided that the density function of the random errors, f_0, is symmetric. This implies that both

$$d(\boldsymbol{\beta}_0) \xrightarrow{P} 0 \quad \text{and} \quad d(\hat{\boldsymbol{\beta}}) \xrightarrow{P} 0\,, \quad \text{as} \quad N \longrightarrow \infty\,.$$

Therefore, the consistency of $\hat{\boldsymbol{\beta}}$ is guaranteed as long as $\mathbb{E}(\mathbf{Z}^{\boldsymbol{\beta}}) = 0$ at, and only at, $\boldsymbol{\beta} = \boldsymbol{\beta}_0$ when the parametric space is compact. □

Theorem 2.3. [**Asymptotic Normality**] Under assumptions **(a2)**-**(a8)** in Appendix A, the central limit property of the QDE $\hat{\boldsymbol{\beta}}$ gives

$$(\mathbf{W}^T\mathbf{W})^{-\frac{1}{2}}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0) \quad \xrightarrow{L} \quad \mathrm{N}\,(\mathbf{0}, \boldsymbol{\Sigma}_2)\,, \tag{10}$$

where $\boldsymbol{\Sigma}_2 = [(\mathbf{W}^T\tilde{\mathbf{X}})^{-1}(\tilde{\mathbf{X}}^T\mathbf{W})^{-1}](\mathbf{S}_0^T\mathbf{Q}\mathbf{S}_0)^{-2}(\mathbf{S}_0^T\mathbf{Q}\boldsymbol{\Sigma}\mathbf{Q}\mathbf{S}_0)$.

Proof. See Appendix A.

Corollary 2.1. The minimum asymptotic variance of the QDE $\hat{\boldsymbol{\beta}}$, $\boldsymbol{\Sigma}_1$, is reached when the weights matrix $\mathbf{W} = \tilde{\mathbf{X}}(\tilde{\mathbf{X}}^T\tilde{\mathbf{X}})^{-1}$ and the $k \times k$ matrix $\mathbf{Q} = \boldsymbol{\Sigma}^{-1}$. That is, $\mathrm{Var}(\hat{\boldsymbol{\beta}}) = (\tilde{\mathbf{X}}^T\tilde{\mathbf{X}})^{-1}(\mathbf{S}_0^T\boldsymbol{\Sigma}^{-1}\mathbf{S}_0)^{-1}$.

Proof. The "optimal weights" can be chosen to minimize (28). Using a generalized Cauchy-Schwartz inequality it is easy to verify that $\mathbf{W} = \tilde{\mathbf{X}}(\tilde{\mathbf{X}}^T\tilde{\mathbf{X}})^{-1}$. Also, if $\boldsymbol{\Sigma}$ is invertible, the optimal choice of $\mathbf{Q}$, in the sense of minimizing the variance covariance matrix $\boldsymbol{\Sigma}_1$, is $\mathbf{Q} = \boldsymbol{\Sigma}^{-1}$. □

2.2 Influence Function of the QD Estimator

Let $\hat{G}_i$ be a degenerate distribution at $\tilde{y}_i$ and define $G_i(\tilde{y}_i) = F_0(\tilde{y}_i - \tilde{\mathbf{x}}_i^T\boldsymbol{\beta}_0)$. Then the QDE, $\hat{\boldsymbol{\beta}}$, can be considered as the statistical functional $\hat{\boldsymbol{\beta}} = \boldsymbol{\beta}(\hat{G}_1, \ldots, \hat{G}_N)$, where $\boldsymbol{\beta}(G_1, \ldots, G_N)$ is defined implicitly as a solution of the $p-$system of equations

$$\frac{\partial}{\partial\boldsymbol{\beta}}[\mathbf{Z}^{\boldsymbol{\beta}}]^T(\mathbf{I}_p \otimes \mathbf{Q})\mathbf{Z}^{\boldsymbol{\beta}} = \mathbf{0}\,,$$

with $\mathbf{Z}^{\boldsymbol{\beta}} = ([\mathbf{Z}_1^{\boldsymbol{\beta}}]^T, \ldots, [\mathbf{Z}_p^{\boldsymbol{\beta}}]^T)^T$, and

$$\mathbf{Z}_j^{\boldsymbol{\beta}} = \left[\sum_{i=1}^{N} \int_{-\infty}^{\infty} w_{ij} h_1(\tilde{y} - \tilde{\mathbf{x}}_i^T \boldsymbol{\beta}) dG_i(\tilde{y}), \ldots, \sum_{i=1}^{N} \int_{-\infty}^{\infty} w_{ij} h_k(\tilde{y} - \tilde{\mathbf{x}}_i^T \boldsymbol{\beta}) dG_i(\tilde{y}) \right]^T .$$

Proposition 2.1. Let $G_{l,\lambda} = (1-\lambda)G_l + \lambda\delta^{\eta_l}$, where δ^{η_l} denotes a degenerate distribution at η_l and $\lambda \in (0,1)$. Let $H(\boldsymbol{\beta}, \lambda) = \frac{\partial}{\partial \boldsymbol{\beta}}[\mathbf{Z}^{\boldsymbol{\beta}}]^T(\mathbf{I}_p \otimes \mathbf{Q})\mathbf{Z}^{\boldsymbol{\beta}}$, if G_l in $\mathbf{Z}^{\boldsymbol{\beta}}$ is replaced by $G_{l,\lambda}$ then the influence function of an observation η_l at $\tilde{x}_l^T$ is given by

$$\mathrm{IF}(\eta_l, \tilde{x}_l^T) = -\left[\frac{\partial H}{\partial \boldsymbol{\beta}}\right]^{-1}\left[\frac{\partial H}{\partial \lambda}\right] ,$$

evaluated at $\boldsymbol{\beta} = \boldsymbol{\beta}_0$ and $\lambda = 0$.

Proof. Under the assumption that $G_{l,\lambda} = (1-\lambda)G_l + \lambda\delta^{\eta_l}$, the influence function of an observation η_l at $\tilde{x}_l^T$ can be written as

$$\mathrm{IF}(\eta_l, \tilde{x}_l^T; \boldsymbol{\beta}, G_{l,\lambda}) = \frac{\partial \boldsymbol{\beta}(G_1, \ldots, G_{l,\lambda}, \ldots, G_N)}{\partial \lambda}\Big|_{\lambda=0} .$$

Now, if G_l in $\mathbf{Z}^{\boldsymbol{\beta}}$ is replaced by $G_{l,\lambda}$, we have that

$$H[\boldsymbol{\beta}(G_1, \ldots, G_{l,\lambda}, \ldots, G_N)] = 0 .$$

Thus

$$\frac{\partial H}{\partial \boldsymbol{\beta}}\Big|_{\substack{\hat{\boldsymbol{\beta}}=\boldsymbol{\beta}(G_1,\ldots,G_{l,\lambda},\ldots,G_N) \\ \lambda=0}} \times \frac{\partial \boldsymbol{\beta}(G_1, \ldots, G_{l,\lambda}, \ldots, G_N)}{\partial \lambda}\Big|_{\lambda=0} + \frac{\partial H}{\partial \lambda}\Big|_{\lambda=0} = 0$$

and the result follows. □

Corollary 2.2. If the conditions of Proposition 2.1 hold, then the vector of influence functions of $\hat{\boldsymbol{\beta}}$ can be expressed as

$$\mathrm{IF}(\eta_l, \tilde{x}_l^T) = (\mathbf{S}_0^T\mathbf{Q}\mathbf{S}_0)^{-1}[(\mathbf{W}^T\tilde{\mathbf{X}})(\tilde{\mathbf{X}}^T\mathbf{W})]^{-1}[\mathbf{W}^T\tilde{\mathbf{X}} \otimes \mathbf{S}_0^T\mathbf{Q}][\mathbf{w}_l^T \otimes h(\eta_l - \mathbf{x}_l^T\boldsymbol{\beta}_0)] .$$

Example 1 (Revisited). We contaminate the data set in Table 1 by adding the observation $\mathbf{x}_{14}^T = (1, 99999)$, $y_{14} = 5$, with a count of $n_{14} = 30$. This observation is an outlier in both y and x, under the logistic regression assumption. The recalculated parameter estimates for the contaminated logistic regression are given for the two methods in Table 3.

	MLE	QDE
$\hat{\beta}_0$	1.770983	1.618106
$\hat{\beta}_1$	-2.960833E-6	7.74599E-6

Table 3: Parameter estimates with outlier

From Tables 2 and 3 we see that the MLE's are greatly affected by the presence of a single outlier, while the QDE remains relatively stable. We use here the %-change as a measure of comparison between parameter estimates.

Table 4 clearly shows that the QDE is more robust to outliers than the MLE.

	MLE	QDE
$\hat{\beta}_0$	7.3%	1.9%
$\hat{\beta}_1$	132.5%	11.9%

Table 4: %-Change due to oulier

3 Robust QD Estimator for the Multinomial Logistic Regression Model

In this section we propose an extension of the QDE to the multinomial logistic regression model. We use ideas similar to those developed in Section 2 for the case of binary responses.

Consider an individual characterized by a vector $\mathbf{x}_i^T = (x_{i1}, \ldots, x_{ip}) \in \mathbb{R}^p$, with p (discrete or continuous) explanatory variables. Let $G_1, \ldots, G_g$ be all the possible groups in which this individual can be classified. We are interested in estimating the probability $\mathbb{P}(Y_i = j|\mathbf{x}_i)$, for $j = 1, \ldots, g$, that an individual with explanatory variable $\mathbf{x}_i$ belong to one of the g groups.

Assume a random sample from populations $G_1, \ldots, G_g$ and denote by N the number of different vectors $\mathbf{x}_i$. Then let n_i be the number of observations at $\mathbf{x}_i$ for $i = 1, \ldots, N$ and y_{ji} the number of G_j-observations at $\mathbf{x}_i$, with $n_i = \sum_{j=1}^{g} y_{ji}$.

Fixing the last classification group G_g and comparing to it the inclusion probabilities of every other group, we say that an observation $\mathbf{x}_i$ satisfies the

logistic assumptions if

$$\ln\left[\frac{\mathbb{P}(Y_i = j|\mathbf{x}_i)}{\mathbb{P}(Y_i = g|\mathbf{x}_i)}\right] = \mathbf{x}_i^T\boldsymbol{\beta}_j \ , \quad j = 1, \ldots, g-1 \ ,$$

or correspondingly,

$$\pi_j(\mathbf{x}_i) = \mathbb{P}(Y_i = j|\mathbf{x}_i) = \frac{\exp\left(\mathbf{x}_i^T\boldsymbol{\beta}_j\right)}{1 + \sum_{l=1}^{g-1}\exp\left(\mathbf{x}_i^T\boldsymbol{\beta}_l\right)} \ , \quad j = 1, \ldots, g \ ,$$

where $\boldsymbol{\beta}_j^T = (\beta_{1j}, \ldots, \beta_{pj})$ is a vector of unknown parameters and $\boldsymbol{\beta}_g = \mathbf{0}$.

Let $\boldsymbol{\beta}^T = (\boldsymbol{\beta}_1^T, \ldots, \boldsymbol{\beta}_{g-1}^T)$ be the $p(g-1)$ dimensional column vector of unknown parameters.

Note that the random variables $(Y_{1i}, \ldots, Y_{gi})$ have a multinomial distribution with n_i trials and cell probabilities $\pi_1(\mathbf{x}_i), \ldots, \pi_g(\mathbf{x}_i)$. Their joint probability mass function is

$$f(y_{1i}, \ldots, y_{gi}) = \frac{n_i!}{y_{1i}! \cdots y_{gi}!}\,\pi_1(\mathbf{x}_i)^{y_{1i}} \cdots \pi_g(\mathbf{x}_i)^{y_{gi}} \ ,$$

with $n_i = \sum_{j=1}^{g} y_{ji}$, for $i = 1, \ldots, N$.

Then the marginal distribution of the Y_{ji} is binomial with index n_i and probability $\pi_j(\mathbf{x}_i)$, that is

$$Y_{ji} \sim \text{Binomial}(n_i, \pi_j(\mathbf{x}_i)) \ , \quad j = 1, \ldots, g \quad ; \quad i = 1, \ldots, N \ .$$

The following is an extension of Definition 2.1 to this multinomial case.

Definition 3.1. For $i = 1, \ldots, N$ fixed, let $(y_{1i}, \ldots, y_{gi})$ be observations from $(Y_{1i}, \ldots, Y_{gi}) \sim \text{Multinomial}[n_i, (\pi_1(\mathbf{x}_i), \ldots, \pi_g(\mathbf{x}_i))]$. Then, the relative frequency P_{ji} is defined, for each $j = 1, \ldots, g$, as

$$P_{ji} = \begin{cases} \frac{1}{2n_i} & \text{if} \quad Y_{ji} = 0 \\ \frac{Y_{ji}}{n_i} & \text{if} \quad 1 \le Y_{ji} \le n_i - 1 \\ 1 - \frac{1}{2n_i} & \text{if} \quad Y_{ji} = n_i \end{cases} \quad , \quad i = 1, \ldots, N \ . \tag{11}$$

Assumptions: Under the above definitions it is assumed that

(a) there exist $\pi_j(\mathbf{x}_i) \in (0,1)$, for $j = 1, \ldots, g$ $i = 1, \ldots, N$, such that if $\min_{1 \le i \le N}(n_i) \longrightarrow \infty$, then

$$(P_{11}, \ldots, P_{gN}) \longrightarrow (\pi_1(\mathbf{x}_1), \ldots, \pi_g(\mathbf{x}_N)) \ , \quad \text{almost surely} \ , \tag{12}$$

(b) there exists exactly one vector $\boldsymbol{\beta}_j^* \in \mathbb{R}^p$, for each $j = 1, \ldots, g$, such that, for all $\boldsymbol{\beta}_j \neq \boldsymbol{\beta}_j^*$

$$\left| \left\{ i; \pi_j(\mathbf{x}_i) = \frac{e^{\mathbf{x}_i^T \boldsymbol{\beta}_j^*}}{\sum_{l=1}^{g} e^{\mathbf{x}_i^T \boldsymbol{\beta}_l^*}} \right\} \right| \geq n_j^* > \left| \left\{ i; \pi_j(\mathbf{x}_i) = \frac{e^{\mathbf{x}_i^T \boldsymbol{\beta}_j}}{\sum_{l=1}^{g} e^{\mathbf{x}_i^T \boldsymbol{\beta}_l}} \right\} \right| \ , \tag{13}$$

where $n_j^* = \lfloor \frac{N}{2} \rfloor + \lfloor \frac{p}{2} \rfloor$, with $\lfloor z \rfloor$ being the largest integer less than or equal to z.

The argument generalizes that given in Theorem 2.1 for a multinomial logistic regression model. Again assume that all values of n_i are reasonably large, such that the results are asymptotic for $n_\cdot = \sum_{i=1}^{N} n_i \to \infty$, where $\frac{n_i}{n_\cdot} \to c_i \in (0,1)$, with N and p remaining fixed.

Theorem 3.1. Consider the above multinomial logistic model and suppose that n_i is large. Then the multinomial logit transform $[\ln(\frac{P_{1i}}{P_{gi}}), \ldots, \ln(\frac{P_{(g-1)i}}{P_{gi}})]^T$ is approximately normally distributed, with mean $\mathbf{x}_i^T \boldsymbol{\beta}$ and variance given by $n_i^{-1}[\pi_1(\mathbf{x}_i)^{-1} + \cdots + (g-1)\pi_g(\mathbf{x}_i)^{-1}]$.

Proof. See Appendix B.

Definition 3.2. (a) Let $\underline{\mathbf{X}}^T = (\underline{\mathbf{X}}_1^T, \ldots, \underline{\mathbf{X}}_N^T)$ be the $N(g-1) \times p(g-1)$ matrix of (discrete or continuous) transformed explanatory variables $X_1, \ldots, X_p$, where

$$\underline{\mathbf{X}}_i^T = (\underline{\mathbf{x}}_{1i}^T, \underline{\mathbf{x}}_{2i}^T, \ldots, \underline{\mathbf{x}}_{(g-1)i}^T) = \begin{pmatrix} v_{1i}\mathbf{x}_i^T & \mathbf{0}^T & \cdots & \mathbf{0}^T \\ \mathbf{0}^T & v_{2i}\mathbf{x}_i^T & \cdots & \mathbf{0}^T \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{0}^T & \mathbf{0}^T & \cdots & v_{(g-1)i}\mathbf{x}_i^T \end{pmatrix} \ ,$$

with $v_{ji} = \{n_i P_{ji}(1 - P_{ji})\}^{\frac{1}{2}}$, for $j = 1, \ldots, g-1$ and $\mathbf{x}_i^T = (x_{i1}, \ldots, x_{ip})$, for $i = 1, \ldots, N$.

(b) Let $\underline{\mathbf{Y}}^T = (\underline{\mathbf{Y}}_1^T, \ldots, \underline{\mathbf{Y}}_N^T)$ be the $N(g-1)\times 1$ vector of multinomial logit transforms, where $\underline{\mathbf{Y}}_i^T = (\underline{Y}_{1i}, \ldots, \underline{Y}_{(g-1)i})$, with $\underline{Y}_{ji} = v_{ji} \ln\left(\frac{P_{ji}}{P_{gi}}\right)$, for $j = 1, \ldots, g-1$ and $i = 1, \ldots, N$.

(c) By means of (a) and (b) we define the "residual" as

$$\underline{r}_{ji} = \underline{y}_{ji} - \underline{\mathbf{x}}_{ji}^T \boldsymbol{\beta} \ , \quad \text{for } j = 1, \ldots, g-1 \quad ; \quad i = 1, \ldots, N \ . \tag{14}$$

In the multinomial logistic model the radom errors are defined by

$$\underline{r}_{ji} = \underline{y}_{ji} - \underline{\mathbf{x}}_{ji}^T \boldsymbol{\beta}_0^* \ , \quad \text{for } j = 1, \ldots, g-1 \quad ; \quad i = 1, \ldots, N \ ,$$

where $\boldsymbol{\beta}_0^* = (\boldsymbol{\beta}_{01}^T, \ldots, \boldsymbol{\beta}_{0(g-1)}^T)$ is the $p(g-1)$ dimensional column vector of unknown parameters. We assume that these errors are independent and identically distributed. Their common distribution function, F_0^*, is unknown but assumed to be absolutely continuous, with a density function f_0^* symmetric around zero. Applying Theorem 3.1 one easily checks that both, the expected value and the index of skewness of these random errors are equal to zero.

Define

$$\hat{F}_t^{\boldsymbol{\beta}}(y) = \sum_{i=1}^{N} \sum_{j=1}^{g-1} w_{jti} I(\underline{y}_{ji} - \underline{\mathbf{x}}_{ji}^T \boldsymbol{\beta} \leq y) \ , \quad \text{for } t = 1, \ldots, p \ , \tag{15}$$

where I denotes an indicator function and w_{jti} are known weights. Similarly, define

$$F_t^0(y) = \sum_{i=1}^{N} \sum_{j=1}^{g-1} w_{jti} F_0^*(y) \ , \quad \text{for } t = 1, \ldots, p \ . \tag{16}$$

Note that the functions $\hat{F}_t^{\boldsymbol{\beta}}$ are empirical processes based on the residuals and weights $w_{jt1}, \ldots, w_{jtN}$, while F_t^0 are the corresponding theoretical distributions.

Now we define for $t = 1, \ldots, p$

$$\begin{aligned}
\mathbf{Z}_t^{\boldsymbol{\beta}} &= \left[\int_{-\infty}^{\infty} h_1(x) d\hat{F}_t^{\boldsymbol{\beta}}(x), \ldots, \int_{-\infty}^{\infty} h_k(x) d\hat{F}_t^{\boldsymbol{\beta}}(x)\right]^T , \\
&= \left[\sum_{i=1}^{N}\sum_{j=1}^{g-1} w_{jti} h_1(\underline{y}_{ji} - \underline{\mathbf{x}}_{ji}^T\boldsymbol{\beta}), \ldots, \sum_{i=1}^{N}\sum_{j=1}^{g-1} w_{jti} h_k(\underline{y}_{ji} - \underline{\mathbf{x}}_{ji}^T\boldsymbol{\beta})\right]^T , \\
\text{and} \quad \mathbf{Z}_t^0 &= \left[\int_{-\infty}^{\infty} h_1(x) dF_t^0(x), \ldots, \int_{-\infty}^{\infty} h_k(x) dF_t^0(x)\right]^T ,
\end{aligned}$$

where $h_1, \ldots, h_k$ is a fixed choice of odd functions, that is $h_l(x) = -h_l(-x)$, for $x \neq 0$ and $h_l(0) = 0$.

The QDE for the multinomial logistic model (QDM) is the vector $\hat{\boldsymbol{\beta}}_M$ which minimizes the following sum of quadratic forms

$$d_M(\boldsymbol{\beta}) = (\mathbf{Z}_1^{\boldsymbol{\beta}} - \mathbf{Z}_1^0)^T \underline{\mathbf{Q}} (\mathbf{Z}_1^{\boldsymbol{\beta}} - \mathbf{Z}_1^0) + \cdots + (\mathbf{Z}_p^{\boldsymbol{\beta}} - \mathbf{Z}_p^0)^T \underline{\mathbf{Q}} (\mathbf{Z}_p^{\boldsymbol{\beta}} - \mathbf{Z}_p^0) , \quad (17)$$

where $\underline{\mathbf{Q}}$ denotes a $k \times k$ constant, symmetric, positive-definite matrix.

Furthermore, since $\mathbf{Z}_t^0 = 0$ for $t = 1, \ldots, p$, when h is odd, minimizing (17) with respect to $\boldsymbol{\beta}$, is reduced to minimizing

$$d_M(\boldsymbol{\beta}) = [\mathbf{Z}_1^{\boldsymbol{\beta}}]^T \underline{\mathbf{Q}} \mathbf{Z}_1^{\boldsymbol{\beta}} + \cdots + [\mathbf{Z}_p^{\boldsymbol{\beta}}]^T \underline{\mathbf{Q}} \mathbf{Z}_p^{\boldsymbol{\beta}} . \quad (18)$$

Calling $\mathbf{Z}^{\boldsymbol{\beta}} = ([\mathbf{Z}_1^{\boldsymbol{\beta}}]^T, \ldots, [\mathbf{Z}_p^{\boldsymbol{\beta}}]^T)^T$ and using Kronecker's product notation, we can express (18) more concisely as

$$d_M(\boldsymbol{\beta}) = [\mathbf{Z}^{\boldsymbol{\beta}}]^T (\mathbf{I}_p \otimes \underline{\mathbf{Q}}) \mathbf{Z}^{\boldsymbol{\beta}} , \quad (19)$$

where $\mathbf{I}_p$ denotes the identity matrix of order p.

The QDM estimator $\hat{\boldsymbol{\beta}}_M$ is the vector which minimizes (19) with respect to $\boldsymbol{\beta}$.

3.1 Asymtotic Properties of the QDM Estimator

In this section we derive the asymptotic properties of the QDM estimator, such as consistency and asymptotic normality.

Definition 3.3. Let $\underline{\mathbf{W}}^T = (\underline{\mathbf{W}}_1^T, \ldots, \underline{\mathbf{W}}_N^T)$ be the $N(g-1) \times p$ matrix of weights used in (15), where

$$\underline{\mathbf{W}}_i^T = \begin{pmatrix} w_{11i} & \cdots & w_{1pi} \\ w_{21i} & \cdots & w_{2pi} \\ \vdots & \ddots & \vdots \\ w_{(g-1)1i} & \cdots & w_{(g-1)pi} \end{pmatrix} , \quad \text{for } i = 1, \ldots, N .$$

Theorem 3.2. [**Consistency**] Consider the matrix of weights $\underline{\mathbf{W}}$ defined above and the $N(g-1) \times p(g-1)$ matrix $\underline{\mathbf{X}}$ given in Definition 3.2. These are assumed to have rank p and $p(g-1)$, respectively. If $\underline{\mathbf{W}}$ satisfies assumption **(*b*1)** given in Appendix B, then the QDM estimator $\hat{\boldsymbol{\beta}}_M$, obtained minimizing the distance $d_M(\boldsymbol{\beta})$, is consistent.

Proof. Chebyshev's inequality and assumption **(*b*1)** give that $\mathbf{Z}^{\boldsymbol{\beta}_0^*} \xrightarrow{P} 0$, provided that the density function of the random errors, f_0^*, is symmetric. This implies that both

$$d_M(\boldsymbol{\beta}_0^*) \xrightarrow{P} 0 \quad \text{and} \quad d_M(\hat{\boldsymbol{\beta}}_M) \xrightarrow{P} 0 , \quad \text{as } N \to \infty .$$

Therefore, the consistency of $\hat{\boldsymbol{\beta}}_M$ is guaranteed as long as $\mathbb{E}(\mathbf{Z}^{\boldsymbol{\beta}}) = 0$ at, and only at $\boldsymbol{\beta} = \boldsymbol{\beta}_0^*$, when the parametric space is compact. □

Theorem 3.3. [**Asymptotic Normality**] Under assumptions **(*b*2)** to **(*b*8)** given in Appendix B, the asymptotic distribution of the QDM estimator $\hat{\boldsymbol{\beta}}_M$ is given by

$$(\hat{\boldsymbol{\beta}}_M - \boldsymbol{\beta}_0^*) \xrightarrow{L} \mathrm{N}(\mathbf{0}, \boldsymbol{\Sigma}_3) , \tag{20}$$

where the variance-covariance matrix $\boldsymbol{\Sigma}_3 = \mathbf{A}_3(\underline{\mathbf{W}}^T\underline{\mathbf{W}})\mathbf{A}_3^T(\underline{\mathbf{S}}_0^T\underline{\mathbf{Q}}\boldsymbol{\Sigma}^*\underline{\mathbf{Q}}\underline{\mathbf{S}}_0)$ and $\mathbf{A}_3 = (\underline{\mathbf{S}}_0^T\underline{\mathbf{Q}}\underline{\mathbf{S}}_0)[(\underline{\mathbf{X}}^T\underline{\mathbf{W}})(\underline{\mathbf{W}}^T\underline{\mathbf{X}})]^{-1}(\underline{\mathbf{X}}^T\underline{\mathbf{W}})$.

Proof. See Appendix B.

Corollary 3.1. The minimum asymptotic variance $\boldsymbol{\Sigma}_3$ of the QDM estimator $\hat{\boldsymbol{\beta}}_M$ is reached when the weights matrix $\underline{\mathbf{W}} = \underline{\mathbf{X}}(\underline{\mathbf{X}}^T\underline{\mathbf{X}})^{-1}$ and the $k \times k$ matrix $\underline{\mathbf{Q}} = [\boldsymbol{\Sigma}^*]^{-1}$. In that case $\mathrm{Var}(\hat{\boldsymbol{\beta}}_M) = (\underline{\mathbf{X}}^T\underline{\mathbf{X}})^{-1}(\underline{\mathbf{S}}_0^T[\boldsymbol{\Sigma}^*]^{-1}\underline{\mathbf{S}}_0)^{-1}$.

Proof. An argument similar to that given for Corollary 2.1, but applied to the variance-covariance matrix $\boldsymbol{\Sigma}_3$ completes the proof. □

Appendices

A Proof of Theorems in Section 2

Proof of Theorem 2.1: Consider the logit transform

$$\ln\Big(\frac{\pi(\mathbf{x}_i)}{1-\pi(\mathbf{x}_i)}\Big) = \mathbf{x}_i^T\boldsymbol{\beta}\ . \tag{21}$$

Under the condition that neither the number of successes nor the number of failures is too small, expression (21) is reasonably estimated by

$$\ln\Big(\frac{Y_i}{n_i - Y_i}\Big) = \ln\left(\frac{\frac{Y_i}{n_i}}{1-\frac{Y_i}{n_i}}\right)\ ,$$

which we call the empirical logit transform.

More generally, if the parametric function of interest is $L[\pi(\mathbf{x}_i)]$, then consider $L\big(\frac{Y_i}{n_i}\big)$. Now, provided that the variation in $\frac{Y_i}{n_i}$ is relatively small we can write

$$L\Big(\frac{Y_i}{n_i}\Big) \approx L[\pi(\mathbf{x}_i)] + \Big(\frac{Y_i}{n_i} - \pi(\mathbf{x}_i)\Big)L'[\pi(\mathbf{x}_i)]\ ,$$

from which it follows that $L\left(\frac{Y_i}{n_i}\right)$ is approximately normally distributed with mean $L[\pi(\mathbf{x}_i)]$ and variance

$$[L'(\pi(\mathbf{x}_i))]^2\mathrm{Var}\Big(\frac{Y_i}{n_i}\Big) = [L'(\pi(\mathbf{x}_i))]^2\frac{\pi(\mathbf{x}_i)(1-\pi(\mathbf{x}_i))}{n_i}\ .$$

Now consider $L(t) = \ln\big(\frac{t}{1-t}\big)$, then

$$\ln\left(\frac{\frac{Y_i}{n_i}}{1-\frac{Y_i}{n_i}}\right) \approx N(\mathbf{x}_i^T\boldsymbol{\beta}, \{n_i\pi(\mathbf{x}_i)[1-\pi(\mathbf{x}_i)]\}^{-1})\ , \quad \text{for } i = 1, \ldots, N\ .$$

□

Assumptions for Asymptotic Properties

(a1) $\lim_{N\to\infty}\sum_{i=1}^{N} w_{ij}^2 = 0$, for each $j = 1, \ldots, p$,

(a2) $\lim_{N\to\infty}\mathbf{W}^T\tilde{\mathbf{X}}$ exists and is invertible,

(a3) $\lim_{N\to\infty}\sum_{i=1}^{N} w_{ij}^2(v_i x_{il})^2 = 0$, for each $j = 1,\ldots,p\,;\, l = 1,\ldots,p$,

(a4) $\lim_{N\to\infty}\sum_{i=1}^{N}\left|w_{ij}v_i x_{il}\right|$ exists for each $j = 1,\ldots,p\,;\, l = 1,\ldots,p$,

(a5) $\dot{h}_i(x) = \frac{\partial}{\partial x}h_i(x)$ is uniformily continuous and $\mathrm{Var}[\dot{h}(\tilde{r})] < \infty$,

(a6) the $v_i x_{ij}$ values belong to a compact set,

(a7) $\max\limits_{1\le i\le N}\{\mathbf{w}_i^T\boldsymbol{\Sigma}\mathbf{w}_i\}$ is bounded for all N,

(a8) $\underline{\lambda}(\mathbf{W}^T\mathbf{W}\otimes\boldsymbol{\Sigma}) \to \infty$ when $N \to \infty$, where $\underline{\lambda}(\mathbf{M})$ represents the smallest eigenvalue of matrix $\mathbf{M}$ and $\boldsymbol{\Sigma}$ is the variance-covariance matrix of $h(\tilde{r}) = [h_1(\tilde{r}),\ldots,h_k(\tilde{r})]^T$.

Proof of Theorem 2.3: It follows from the form of the asymptotic variance-covariance matrix of $\hat{\boldsymbol{\beta}}$ and the multivariate central limit theorem.

Let $\mathbf{S}_0^T = [\mathbb{E}(\dot{h}_1(\tilde{r})),\ldots,\mathbb{E}(\dot{h}_k(\tilde{r}))]$, where $\dot{h}_i(x) = \frac{\partial}{\partial x}h_i(x)$ and assume that the function d, given by (9), is differentiable. Then $\hat{\boldsymbol{\beta}}$ satisfies the following p-system of equations

$$\frac{\partial}{\partial\boldsymbol{\beta}}[\mathbf{Z}^{\hat{\boldsymbol{\beta}}}]^T(\mathbf{I}_p\otimes\mathbf{Q})\mathbf{Z}^{\hat{\boldsymbol{\beta}}} = \mathbf{0}\,. \tag{22}$$

Under assumptions **(a3)** to **(a6)** and using the properties of Kronecker's product:

$$\begin{aligned}
\frac{\partial}{\partial\boldsymbol{\beta}}\mathbf{Z}^{\hat{\boldsymbol{\beta}}} &= \frac{\partial}{\partial\boldsymbol{\beta}}\mathbf{Z}^{\boldsymbol{\beta}_0} + o_p(1)\,, &(23)\\
\frac{\partial}{\partial\boldsymbol{\beta}}\mathbf{Z}^{\boldsymbol{\beta}_0} &= -\mathbf{W}^T\tilde{\mathbf{X}}\otimes\mathbf{S}_0 + o_p(1)\,, &(24)\\
\frac{\partial}{\partial\boldsymbol{\beta}}[\mathbf{Z}^{\boldsymbol{\beta}_0}]^T(\mathbf{I}\otimes\mathbf{Q})\frac{\partial}{\partial\boldsymbol{\beta}}\mathbf{Z}^{\boldsymbol{\beta}_0} &= (\tilde{\mathbf{X}}^T\mathbf{W})(\mathbf{W}^T\tilde{\mathbf{X}})\otimes(\mathbf{S}_0^T\mathbf{Q}\mathbf{S}_0) + o_p(1)\\
&= (\tilde{\mathbf{X}}^T\mathbf{W})(\mathbf{W}^T\tilde{\mathbf{X}})(\mathbf{S}_0^T\mathbf{Q}\mathbf{S}_0) + o_p(1)\,, &(25)
\end{aligned}$$

where $o_p(1)$ stands for a random infinitesimal term converging in probability.

Substitute (24) and (25) in (22) and use a Taylor's expansion to get

$$(\mathbf{S}_0^T\mathbf{Q}\mathbf{S}_0)(\tilde{\mathbf{X}}^T\mathbf{W})(\mathbf{W}^T\tilde{\mathbf{X}})(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta}_0) = -(\tilde{\mathbf{X}}^T\mathbf{W}\otimes\mathbf{S}_0^T)(\mathbf{I}\otimes\mathbf{Q})\mathbf{Z}^{\boldsymbol{\beta}_0} + o_p(1)\,. \tag{26}$$

Since $\mathbf{Z}^{\boldsymbol{\beta}_0}$ is a vector of sums of independent variables, then under assumptions $(\boldsymbol{a7})$, $(\boldsymbol{a8})$ and the multivariate central limit theorem:

$$(\mathbf{W}^T\mathbf{W} \otimes \boldsymbol{\Sigma})^{-\frac{1}{2}}\mathbf{Z}^{\boldsymbol{\beta}_0} \xrightarrow{L} \mathrm{N}(\mathbf{0}, \mathbf{I}) \; . \tag{27}$$

Using (27) and (26), we have that

$$\mathrm{Var}[-(\tilde{\mathbf{X}}^T\mathbf{W}\otimes\mathbf{S}_0^T)(\mathbf{I}\otimes\mathbf{Q})\mathbf{Z}^{\boldsymbol{\beta}_0}] = (\tilde{\mathbf{X}}^T\mathbf{W}\otimes\mathbf{S}_0^T)(\mathbf{W}^T\mathbf{W}\otimes\mathbf{Q}\boldsymbol{\Sigma}\mathbf{Q})(\mathbf{W}^T\tilde{\mathbf{X}}\otimes\mathbf{S}_0) \; .$$

Then $(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0)$ is asymptotically normal with asymptotic variance-covariance matrix

$$\boldsymbol{\Sigma}_1 = \mathbf{A}(\mathbf{W}^T\mathbf{W} \otimes \mathbf{Q}\boldsymbol{\Sigma}\mathbf{Q})\mathbf{A}^T \; ,$$

where $\mathbf{A} = [(\mathbf{S}_0^T\mathbf{Q}\mathbf{S}_0)(\tilde{\mathbf{X}}^T\mathbf{W})(\mathbf{W}^T\tilde{\mathbf{X}})]^{-1}[\tilde{\mathbf{X}}^T\mathbf{W} \otimes \mathbf{S}_0^T]$.

Finally $\boldsymbol{\Sigma}_1$ can be expressed as

$$\boldsymbol{\Sigma}_1 = (\mathbf{W}^T\tilde{\mathbf{X}})^{-1}(\mathbf{W}^T\mathbf{W})(\tilde{\mathbf{X}}^T\mathbf{W})^{-1}(\mathbf{S}_0^T\mathbf{Q}\mathbf{S}_0)^{-2}(\mathbf{S}_0^T\mathbf{Q}\boldsymbol{\Sigma}\mathbf{Q}\mathbf{S}_0) \; , \tag{28}$$

or equivalently,

$$(\mathbf{W}^T\mathbf{W})^{-\frac{1}{2}}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0) \xrightarrow{L} \mathrm{N}(\mathbf{0}, \boldsymbol{\Sigma}_2) \; ,$$

where $\boldsymbol{\Sigma}_2 = (\mathbf{W}^T\tilde{\mathbf{X}})^{-1}(\tilde{\mathbf{X}}^T\mathbf{W})^{-1}(\mathbf{S}_0^T\mathbf{Q}\mathbf{S}_0)^{-2}(\mathbf{S}_0^T\mathbf{Q}\boldsymbol{\Sigma}\mathbf{Q}\mathbf{S}_0)$. □

B Proof of Theorems in Section 3

Proof of Theorem 3.1: Consider the multinomial logistic model link function

$$l[\boldsymbol{\pi}(\mathbf{x_i})] = \Big[\ln\Big(\frac{\pi_1(\mathbf{x}_i)}{\pi_g(\mathbf{x}_i)}\Big), \ldots, \ln\Big(\frac{\pi_{g-1}(\mathbf{x}_i)}{\pi_g(\mathbf{x}_i)}\Big)\Big]^T = \mathbf{x}_i^T\boldsymbol{\beta} \; , \quad i = 1, \ldots, N \; , \tag{29}$$

where $\boldsymbol{\pi}(\mathbf{x_i}) = (\pi_1(\mathbf{x}_i), \ldots, \pi_g(\mathbf{x}_i))$.

Suppose that $\pi_j(\mathbf{x}_i)$ and $\pi_g(\mathbf{x}_i)$ are never "too small". Then a reasonable estimate of (29) is $l(\frac{Y_{1i}}{n_i}, \ldots, \frac{Y_{gi}}{n_i}) = \Big[\ln\Big(\frac{Y_{1i}}{Y_{gi}}\Big), \ldots, \ln\Big(\frac{Y_{1i}}{Y_{gi}}\Big)\Big]^T$, which we call the empirical logit transform. In general if the function of interest is $l[\boldsymbol{\pi}(\mathbf{x_i})]$, then consider $l(\frac{Y_{1i}}{n_i}, \ldots, \frac{Y_{gi}}{n_i})$ as an estimator.

Now suppose that the variation in $\frac{Y_{ji}}{n_i}$ is relatively small and consider a Taylor's series expansion

$$l\Big(\frac{Y_{1i}}{n_i},\ldots,\frac{Y_{gi}}{n_i}\Big) \approx l[\boldsymbol{\pi}(\mathbf{x_i})] + \sum_{i=1}^{g} \dot{l}_j[\boldsymbol{\pi}(\mathbf{x_i})]\Big(\frac{Y_{ji}}{n_i} - \pi_j(\mathbf{x}_i)\Big)\,, \qquad (30)$$

where $\dot{l}_j[\boldsymbol{\pi}(\mathbf{x_i})] = \frac{\partial}{\partial z_j} l(z_1,\ldots,z_g)\Big|_{z_1=\pi_1(\mathbf{x}_i),\ldots,z_g=\pi_g(\mathbf{x}_i)}$.

Using (30) it can be seen that $l\Big(\frac{Y_{1i}}{n_i},\ldots,\frac{Y_{gi}}{n_i}\Big)$ is approximately normally distributed with mean $l[\boldsymbol{\pi}(\mathbf{x_i})] = \mathbf{x}_i^T\boldsymbol{\beta}$ and variance

$$\begin{aligned}
\mathrm{Var}\left[l\Big(\frac{Y_{1i}}{n_i},\ldots,\frac{Y_{gi}}{n_i}\Big)\right] &= \sum_{j=1}^{g}\{\dot{l}_j[\boldsymbol{\pi}(\mathbf{x_i})]\}^2\mathrm{Var}\Big(\frac{Y_{ji}}{n_i}\Big) \\
&\quad +2\sum_{j>j^*}\dot{l}_j[\boldsymbol{\pi}(\mathbf{x_i})]\dot{l}_{j^*}[\boldsymbol{\pi}(\mathbf{x_i})]\mathrm{Cov}\Big(\frac{Y_{ji}}{n_i},\frac{Y_{j^*i}}{n_i}\Big)\,, \\
&= n_i^{-1}[\pi_1(\mathbf{x}_i)^{-1}+\cdots+\pi_{g-1}(\mathbf{x}_i)^{-1}+(g-1)\pi_g(\mathbf{x}_i)^{-1}]\,.
\end{aligned}$$

□

Assumptions for Asymptotic Properties

(b1) $\lim_{N\to\infty}\sum_{i=1}^{N}\sum_{j=1}^{g-1} w_{jti}^2 = 0$, for each $t = 1,\ldots,p$,

(b2) $\lim_{N\to\infty}(\underline{\mathbf{X}}^T\underline{\mathbf{W}})(\underline{\mathbf{W}}^T\underline{\mathbf{X}})$ exists and is invertible,

(b3) $\lim_{N\to\infty}\sum_{i=1}^{N}\sum_{j=1}^{g-1} w_{jti}^2[v_{ji}x_{ti}]^2 = 0$, for each $t = 1,\ldots,p$,

(b4) $\lim_{N\to\infty}\sum_{i=1}^{N}\sum_{j=1}^{g-1}\big|w_{jti}v_{ji}x_{ti}\big|$ exists for each $t = 1,\ldots,p$,

(b5) $\dot{h}_i(x) = \frac{\partial}{\partial x}h_i(x)$ is uniformily continuous and $\mathrm{Var}[\dot{h}(\underline{r})] < \infty$,

(b6) the $v_{ji}x_{ti}$ values belong to a compact set,

(b7) $\max_{1\le i\le N}\{\mathbf{w}_i^T\boldsymbol{\Sigma}^*\mathbf{w}_i\}$ is bounded for all N,

(b8) $\underline{\lambda}(\underline{\mathbf{W}}^T\underline{\mathbf{W}}\otimes\boldsymbol{\Sigma}^*)\to\infty$ if $N\to\infty$, where $\underline{\lambda}(\mathbf{M})$ represents the smallest eigenvalue of matrix $\mathbf{M}$ and $\boldsymbol{\Sigma}^*$ is the variance-covariance matrix of $h(\underline{r}) = [h_1(\underline{r}),\ldots,h_k(\underline{r})]^T$.

Proof of Theorem 3.3: It follows from the form of the asymptotic variance-covariance matrix of $\hat{\boldsymbol{\beta}}_M$ and the multivariate central limit theorem.

Consider $\underline{\mathbf{S}}_0^T = [\mathbb{E}(\dot{h}_1(\underline{r})), \ldots, \mathbb{E}(\dot{h}_k(\underline{r}))]$, where $\dot{h}_i(x) = \frac{\partial}{\partial x} h_i(x)$ and assume that the function d_M, given by (19), is differentiable. Then $\hat{\boldsymbol{\beta}}_M$ satisfies the following $p(g-1)$-system of equations

$$\frac{\partial}{\partial \boldsymbol{\beta}} [\mathbf{Z}^{\hat{\boldsymbol{\beta}}_M}]^T (\mathbf{I}_p \otimes \mathbf{Q}) \mathbf{Z}^{\hat{\boldsymbol{\beta}}_M} = \mathbf{0} \; . \tag{31}$$

From assumptions $(\boldsymbol{b3})$ to $(\boldsymbol{b6})$ and the properties of Kronecker's product:

$$\begin{aligned}
\frac{\partial}{\partial \boldsymbol{\beta}} \mathbf{Z}^{\hat{\boldsymbol{\beta}}_M} &= \frac{\partial}{\partial \boldsymbol{\beta}} \mathbf{Z}^{\boldsymbol{\beta}_0^*} + o_p(1) \; , & (32) \\
\frac{\partial}{\partial \boldsymbol{\beta}} \mathbf{Z}^{\boldsymbol{\beta}_0^*} &= -\underline{\mathbf{W}}^T \underline{\mathbf{X}} \otimes \underline{\mathbf{S}}_0 + o_p(1) \; , & (33) \\
\frac{\partial}{\partial \boldsymbol{\beta}} [\mathbf{Z}^{\boldsymbol{\beta}_0^*}]^T (\mathbf{I} \otimes \underline{\mathbf{Q}}) \frac{\partial}{\partial \boldsymbol{\beta}} \mathbf{Z}^{\boldsymbol{\beta}_0^*} &= (\underline{\mathbf{X}}^T \underline{\mathbf{W}} \otimes \underline{\mathbf{S}}_0^T)(\mathbf{I} \otimes \underline{\mathbf{Q}})(\underline{\mathbf{W}}^T \underline{\mathbf{X}} \otimes \underline{\mathbf{S}}_0^T) + o_p(1) \\
&= (\underline{\mathbf{X}}^T \underline{\mathbf{W}})(\underline{\mathbf{W}}^T \underline{\mathbf{X}})(\underline{\mathbf{S}}_0^T \underline{\mathbf{Q}} \underline{\mathbf{S}}_0) + o_p(1) \; , & (34)
\end{aligned}$$

where $o_p(1)$ stands for a random infinitesimal term term converging in probability.

Substite (33) and (34) in (31) and use a Taylor's expansion to get

$$(\underline{\mathbf{S}}_0^T \underline{\mathbf{Q}} \underline{\mathbf{S}}_0)(\underline{\mathbf{X}}^T \underline{\mathbf{W}})(\underline{\mathbf{W}}^T \underline{\mathbf{X}})(\hat{\boldsymbol{\beta}}_M - \boldsymbol{\beta}_0^*) = -(\underline{\mathbf{X}}^T \underline{\mathbf{W}} \otimes \underline{\mathbf{S}}_0^T)(\mathbf{I} \otimes \underline{\mathbf{Q}}) \mathbf{Z}^{\boldsymbol{\beta}_0^*} + o_p(1) \; . \tag{35}$$

Since $\mathbf{Z}^{\boldsymbol{\beta}_0^*}$ is a vector of sums of independent variables, then under assumptions $(\boldsymbol{b7})$, $(\boldsymbol{b8})$ and the multivariate central limit theorem:

$$\mathbf{Z}^{\boldsymbol{\beta}_0^*} \xrightarrow{L} \mathrm{N}(\mathbf{0}, \underline{\mathbf{W}}^T \underline{\mathbf{W}} \otimes \boldsymbol{\Sigma}^*) \; . \tag{36}$$

From (36) and (35), we obtain that

$$\mathrm{Var}[(\underline{\mathbf{X}}^T \underline{\mathbf{W}} \otimes \underline{\mathbf{S}}_0^T)(\mathbf{I} \otimes \underline{\mathbf{Q}}) \mathbf{Z}^{\boldsymbol{\beta}_0^*}] = (\underline{\mathbf{X}}^T \underline{\mathbf{W}} \otimes \underline{\mathbf{S}}_0^T)(\underline{\mathbf{W}}^T \underline{\mathbf{W}} \otimes \underline{\mathbf{Q}} \boldsymbol{\Sigma}^* \underline{\mathbf{Q}})(\underline{\mathbf{W}}^T \underline{\mathbf{X}} \otimes \underline{\mathbf{S}}_0^T) \; .$$

Thus

$$\mathrm{Var}(\hat{\boldsymbol{\beta}}_M) = \mathbf{A}_2 (\underline{\mathbf{W}}^T \underline{\mathbf{W}} \otimes \underline{\mathbf{Q}} \boldsymbol{\Sigma}^* \underline{\mathbf{Q}}) \mathbf{A}_2^T \; ,$$

where $\mathbf{A}_2 = (\underline{\mathbf{S}}_0^T \underline{\mathbf{Q}} \underline{\mathbf{S}}_0)[(\underline{\mathbf{X}}^T \underline{\mathbf{W}})(\underline{\mathbf{W}}^T \underline{\mathbf{X}})]^{-1} (\underline{\mathbf{X}}^T \underline{\mathbf{W}} \otimes \underline{\mathbf{S}}_0^T)$, or equivalently,

$$\boldsymbol{\Sigma}_3 = \mathbf{A}_3 (\underline{\mathbf{W}}^T \underline{\mathbf{W}}) \mathbf{A}_3^T (\underline{\mathbf{S}}_0^T \underline{\mathbf{Q}} \boldsymbol{\Sigma}^* \underline{\mathbf{Q}} \underline{\mathbf{S}}_0) \; ,$$

where $\mathbf{A}_3 = (\underline{\mathbf{S}}_0^T \underline{\mathbf{Q}} \underline{\mathbf{S}}_0)[(\underline{\mathbf{X}}^T \underline{\mathbf{W}})(\underline{\mathbf{W}}^T \underline{\mathbf{X}})]^{-1} (\underline{\mathbf{X}}^T \underline{\mathbf{W}})$.

Therefore $(\hat{\boldsymbol{\beta}}_M - \boldsymbol{\beta}_0^*)$ is asymptotically normal with asymptotic variance-covariance matrix $\boldsymbol{\Sigma}_3$. □

Acknowledgements

Esteban Flores thanks the University of Talca as well as the joint Ph.D. committee of the Casualty Actuarial Society and the Society of Actuaries for their financial support.

José Garrido is also grateful to the Natural Sciences and Engineering Council of Canada (NSERC) for its financial support through operating grant OGP0036860.

References

[1] Christmann, A. (1994), "Least median of weighted squares in logistic regression with large strata". *Biometrika*, **81**, 413-417.

[2] Graham, A. (1981), *Kronecker Products and Matrix Calculus: with Applications*. Halsted Press, Chichester.

[3] Luong, A. and Thompson, M.E. (1987), "Minimum distance methods based on quadratic distances for transforms". *Canadian Journal of Statistics*, **15**, 239-251.

[4] Luong, A. (1991), " Minimum distance methods based on quadratic distances for transforms in simple linear regression models". *Journal Royal Statistical Society B*, **53**, 465-471.

[5] Luong, A. and Garrido, J. (1992), "Nonparametric estimation based on minimum quadratic distances for the multiple linear regression model". *Cuadernos Aragoneses de Economia*, **2**, 69-78, (in Spanish).